

ABSTRACT

A near-duplicate component includes a fingerprint creation component and a similarity detection component. The fingerprint creation component receives a document of arbitrary size and generates a compact "fingerprint" that describes the contents of the document. The similarity detection component compares multiple fingerprints based on the hamming distance between the fingerprints. When the hamming distance is below a threshold, the documents can be said to be near-duplicates of one another.